Computerized Text Analysis: Classwork 1

Kenneth Benoit

The objective of this class exercise is to get a feel for the workflow of preparing texts for computerized analysis. While this seems simple, in practice texts come in a variety of formats and may need to be converted before they can be processed and analyzed. These formatting issues include the *file type*—such as Microsoft Word documents, Adobe's pdf format, HTML, and various forms of "plain text"—and the *encoding* of files, especially text files, an issue that will affect non-English texts and any texts with extended characters (such as accented letters, bullet point, "smart quotes", etc.).

The other objective is to familiarize yourselves with a workflow for working with texts, and to learn some useful tips for organizing files, renaming files, developing a sequence for working with texts, and so on.

Finally, at a very basic level, the objective is to get our hands dirty start working with actual texts!

Instructions:

1. Download the ten French political documents from `http://www.kenbenoit.net/files/frenchtexts.zip`. These are in a zipped archive, so you will need to unzip them.
2. Try to locate them in a command window or a shell: "command" in Windows or "Terminal.app" in Mac OS/X. You ought to familiarize yourself with the `pwd` and `chdir` or `cd` commands.
3. Examine the texts and their formats. There are a variety of ways to do this, including using your web browser, a text editor, `less` from the command line, `cat <filename> | less`, `file *` if you have a UNIX-based OS (e.g. Mac OS X or Linux).
4. Read up on text encoding issues. Here is a good start: `http://en.wikipedia.org/wiki/Character_encoding`.
5. Convert the files to plain text.
   - To convert Microsoft Office documents to text: "Save As: Text" from Microsoft Word. OpenOffice is also particularly good for this, and it also has the advantage of being multi-platform and free.
   - To convert HTML documents to text: You can use Word or OpenOffice, or you can can open the files in a browser (e.g. Firefox) and "Save As: Text".
   - To convert pdf documents to text, you can use a great utility called `pdftotext`:
     - Windows binaries (and Linux packages): `http://www.foolabs.com/xpdf/download.html`.
     - Mac OS/X binaries: `http://www.bluem.net/downloads/pdftotext_en/`
     - You will have to locate the pdftotext file in the same directory as your texts and then run the program from the command line (or put the binary in a directory specified in your PATH environment variable).
     - You can find the necessary commands by typing: `pdftotext -help` or checking the README file.
6. Inspect and possibly "clean" the texts using a text editor. Be careful with the encodings!